Data, Models, & Concepts for Quantitative Finance

Global Association of Risk Professionals

Webinar – August 2013

Joe Pimbley

http://www.maxwell-consulting.com/

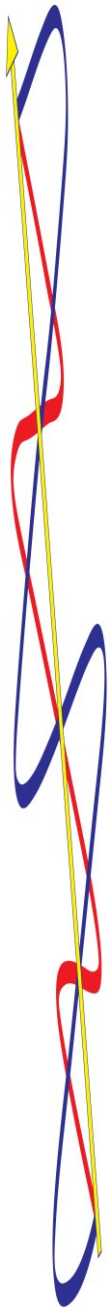
Outline

- Time Series Data Analysis
- Adding Models to the Data
- Adding Concepts to the Model
- Ultimate Result: Necessary but not sufficient

Use: Build Model for Treasury Yield Change Probability

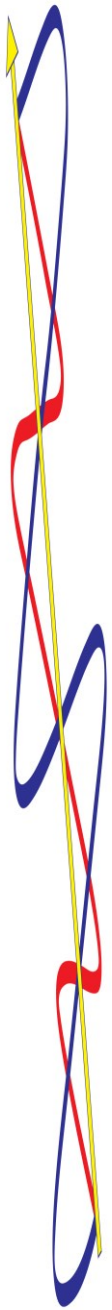
Use: Build Model for Sovereign Default Probability

Use: Build Model for Spread Widening Probability



"A theory is something nobody believes except the person proposing the theory ...."

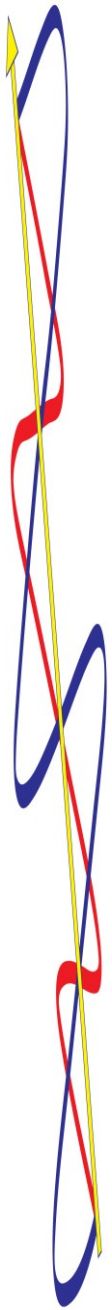
- Albert Einstein



"A theory is something nobody believes except the person proposing the theory ...."

.... whereas an experiment is something everybody believes except the person performing the experiment."

- Albert Einstein



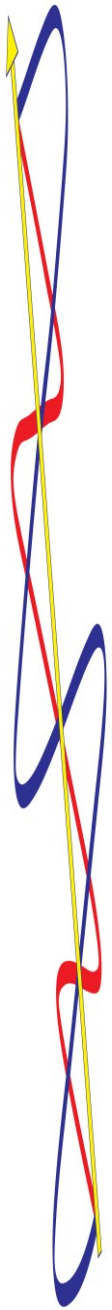
"A theory is something nobody believes except the person proposing the theory ....

.... whereas an experiment is something everybody believes except the person performing the experiment."

- Albert Einstein

Very True!

Corollary for the Financial World: Do not believe that models or data determine probabilities for future market behavior. Rather, it is merely USEFUL AND IMPERATIVE TO KNOW what models and data indicate.



Stochastic Models for Time Evolution of Market Variables begin with Normal and Log-Normal forms

Why?

Math is straightforward (as these things go)

Consistent with market efficiency

Central Limit Theorem

Good place to start – can explain results

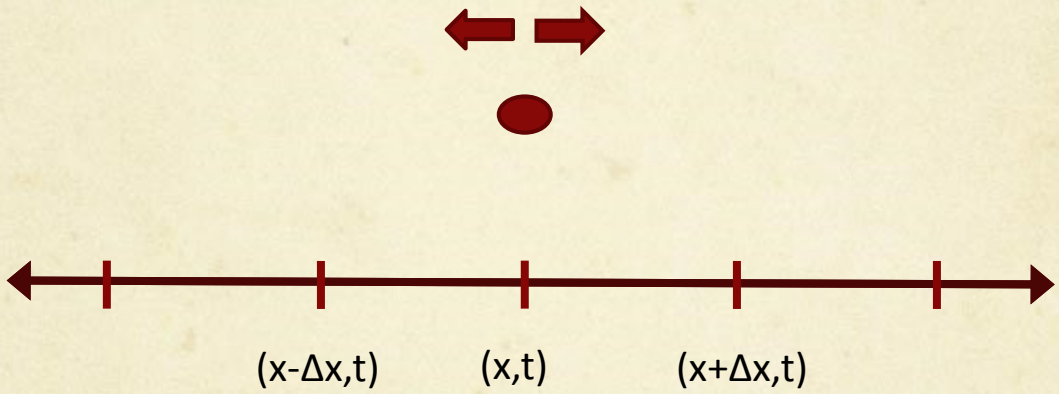
Normal Stochastic Model for yield y with standard deviation s, random variable $\boldsymbol{\varepsilon} \rightarrow N(0,1)$

$$\Delta y = r\,\Delta t + s\,\varepsilon\,\sqrt{\Delta t}$$

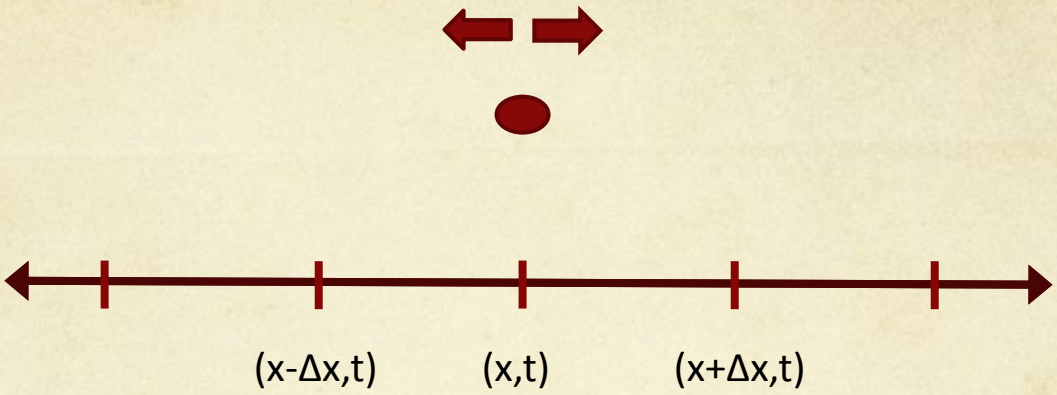
Log-Normal Stochastic Model for spread S with volatility $\boldsymbol{\sigma}$, random variable $\boldsymbol{\varepsilon} \rightarrow N(0,1)$

$$\Delta S = \mu\,S\,\Delta t + \sigma\,S\,\varepsilon\,\sqrt{\Delta t}$$

To elaborate, why does random evolution in time imply the Normal or Log-Normal distributions?



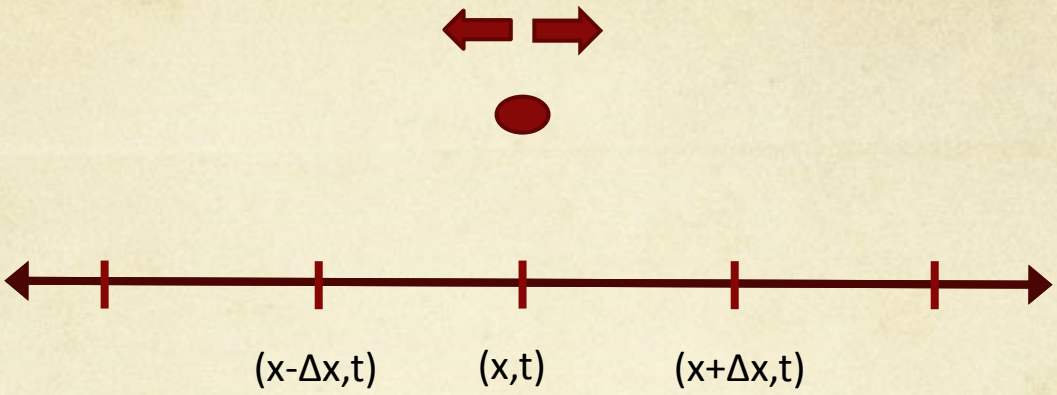
Particle moves randomly with equal probability **α** in either direction during time step **Δt**



$$p(x, t + \Delta t)$$

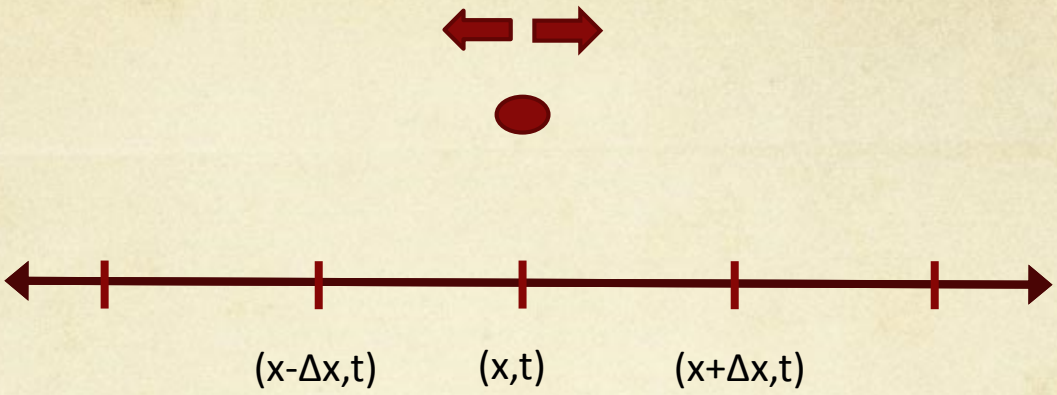
$$= \alpha\, p(x - \Delta x, t) + \alpha\, p(x + \Delta x, t) + (1 - 2\alpha)\, p(x, t)$$

Let $p(x,t)$ be the probability that the particle is at position x at time t



$$\frac{\partial p}{\partial t} = \alpha \frac{(\Delta x)^2}{\Delta t} \frac{\partial^2 p}{\partial x^2} = D \frac{\partial^2 p}{\partial x^2}$$

Applying Taylor series, we get this partial differential equation (PDE) for "diffusion"



$$p(x,t) = \frac{1}{2\sqrt{\pi D t}} \exp\left[\frac{-x^2}{4Dt}\right]$$

Solution to the PDE for a particle that begins at x = 0

It's a normal distribution!

The mean position is zero and the standard deviation is $\sqrt{2Dt}$

The annualized standard deviation of the Δ column values is 1.1%.

Basic method to determine standard deviation for Normal Stochastic Process assumption

		Δ	
date	us10year	us10year	
01/02/62	4.06%	-0.03%	1.1%
01/03/62	4.03%	-0.04%	
01/04/62	3.99%	0.03%	
01/05/62	4.02%	0.01%	
01/08/62	4.03%	0.02%	
01/09/62	4.05%	0.02%	
01/10/62	4.07%	0.01%	
01/11/62	4.08%	0.00%	
01/12/62	4.08%	0.02%	
01/15/62	4.10%	0.03%	
01/16/62	4.13%	-0.01%	
01/17/62	4.12%	-0.01%	
01/18/62	4.11%	0.00%	
01/19/62	4.11%	-0.02%	

NOT these values ....

Do NOT simply take the standard deviation of the "us10year" column! That's not what we want.

		Δ	
date	us10year	us10year	
01/02/62	4.06%	-0.03%	1.1%
01/03/62	4.03%	-0.04%	
01/04/62	3.99%	0.03%	
01/05/62	4.02%	0.01%	
01/08/62	4.03%	0.02%	
01/09/62	4.05%	0.02%	
01/10/62	4.07%	0.01%	
01/11/62	4.08%	0.00%	
01/12/62	4.08%	0.02%	
01/15/62	4.10%	0.03%	
01/16/62	4.13%	-0.01%	
01/17/62	4.12%	-0.01%	
01/18/62	4.11%	0.00%	
01/19/62	4.11%	-0.02%	

The annualized standard deviation of the Δ column values is 18.4%. This is "volatility" for Log-Normal.

Basic method to determine volatility for Log-Normal Stochastic Process assumption

date	Log (us10year)	Δ Log (us10year)	
01/02/62	-3.204	-0.007	18.4%
01/03/62	-3.211	-0.010	
01/04/62	-3.221	0.007	
01/05/62	-3.214	0.002	
01/08/62	-3.211	0.005	
01/09/62	-3.206	0.005	
01/10/62	-3.202	0.002	
01/11/62	-3.199	0.000	
01/12/62	-3.199	0.005	
01/15/62	-3.194	0.007	
01/16/62	-3.187	-0.002	
01/17/62	-3.189	-0.002	
01/18/62	-3.192	0.000	
01/19/62	-3.192	-0.005	

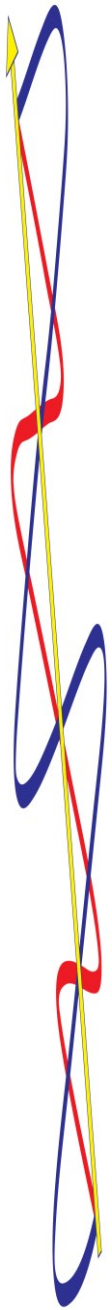
Thus, a Normal Stochastic Model for the 10-year UST yield *y* might use standard deviation ***s = 1.1%*** and the random variable **ε** → N(0,1)

$$\Delta y = r\,\Delta t + s\,\varepsilon\,\sqrt{\Delta t}$$

A Log-Normal Stochastic Model for the 10-year UST yield *y* might use volatility ***σ = 18.4%*** and the random variable **ε** → N(0,1)

$$\Delta y = \mu\, y\,\Delta t + \sigma\, y\,\varepsilon\,\sqrt{\Delta t}$$

(In either case, it's customary to specify / determine the drift terms by a different method.)



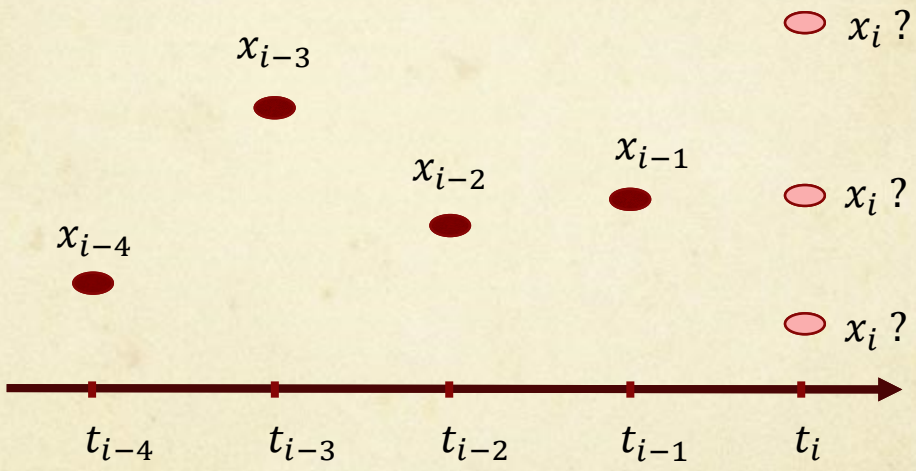
"And now for something .... completely different."

- Monty Python

Auto-Regressive (AR) Estimation

Auto-Regressive (AR) Estimation

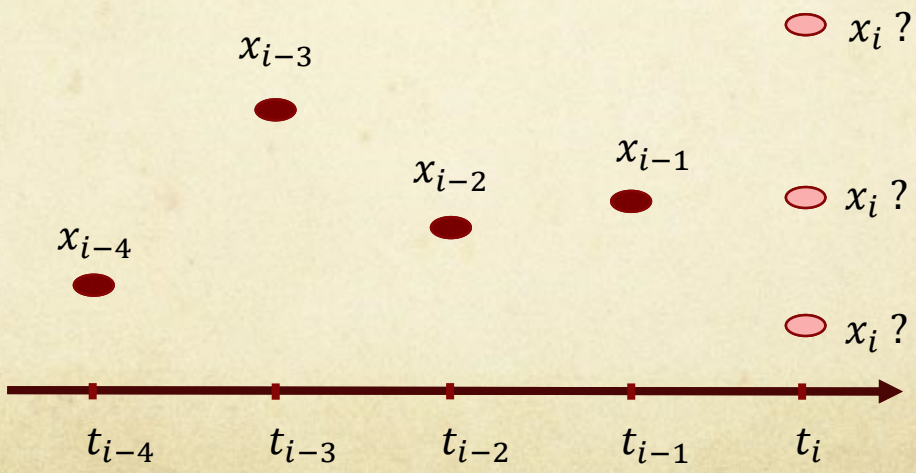
How to predict x_i from prior values of x?



How to predict x_i from prior values of x?

Let the prediction be a linear combination of the prior values:

$$\hat{x}_i = \alpha + \sum_{j=1}^{p} a_j \, x_{i-j}$$

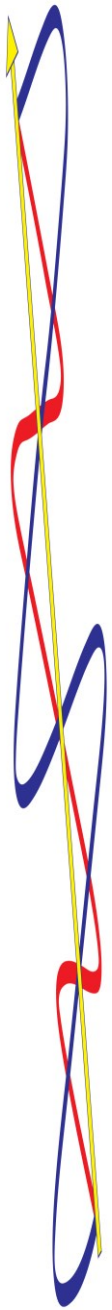


How to predict x_i from prior values of x?

Let the prediction be a linear combination of the prior values:

$$\hat{x}_i = \alpha + \sum_{j=1}^{p} a_j \, x_{i-j}$$

The α and the a_j are constant values that one determines to provide the best predictions for the x_i. The integer p is the "order" of the "AR Model" – designated AR(p).



Auto-Regressive Estimation

Is there any concept behind this AR method? Engineers call it a "Black Box" technique (and use it!).

We find it useful for adding context to stochastic models.

Let the prediction be a linear combination of the prior values:

$$\hat{x}_i = \alpha + \sum_{j=1}^{p} a_j \, x_{i-j}$$

Write the sum of "squared errors":

$$SSE = \sum_{i=p+1}^{N} (x_i - \hat{x}_i)^2$$

$$= \sum_{i=p+1}^{N} \left(x_i - \alpha - \sum_{j=1}^{p} a_j \, x_{i-j} \right)^2$$

Find the α and the a_j by minimizing the sum of squared errors:

$$\frac{\partial\, SSE}{\partial \alpha} = 0$$

$$\frac{\partial\, SSE}{\partial a_j} = 0\,, \qquad \mathrm{j} = 1, \cdots, p$$

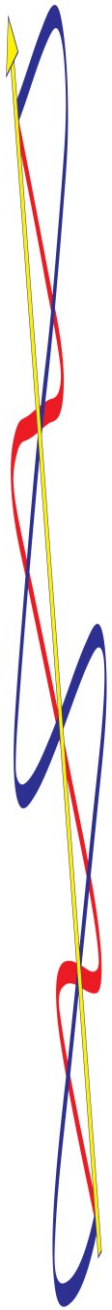
The equations simplify to the form

$$\sum_{j=1}^{p} \Gamma_{mj}\, a_j \;=\; \Gamma_{m0}\,, \qquad m = 1, \cdots, p$$

The elements Γ_{mj} of the "Γ matrix" are known from the historical data values:

$$\Gamma_{mj} \;=\; \sum_{i=p+1}^{N} (x_{i-m} - \mu_m)\left(x_{i-j} - \mu_j\right)\,, \qquad m, j = 0, \cdots, p$$

with $\mu_j \;=\; \sum_{l=p+1}^{N} x_{l-j}/(N-p)\;\;,\;\; j = 0, \cdots, p$



Worthwhile points to make

- Next step is to solve a linear system
- May not need to write all this software yourself, but it helps to be aware of nuances (especially with a small number of data points)
- User chooses the order p – which raises further questions

date	Log (us10year)
01/02/62	-3.204
01/03/62	-3.211
01/04/62	-3.221
01/05/62	-3.214
01/08/62	-3.211
01/09/62	-3.206
01/10/62	-3.202
01/11/62	-3.199
01/12/62	-3.199
01/15/62	-3.194
01/16/62	-3.187
01/17/62	-3.189
01/18/62	-3.192
01/19/62	-3.192

AR Model Output: the RMSE (annualized) is the volatility we found earlier

(0.00)	alpha	**Data Points:**	12763
1.00	AR(1)	**RMSE:**	18.4%
		Mean Error:	(0.00)
		Correlation:	0.05
		Akaike:	-1.01E+05

p = 1

AR Model Output: Notice different results for different *p*

(0.00)	alpha	**Data Points:**	12763
1.00	AR(1)	**RMSE:**	18.4%
		Mean Error:	(0.00)
		Correlation:	0.05
		Akaike:	-1.01E+05

p = 1

(0.00)	alpha	**Data Points:**	12763
1.05	AR(1)	**RMSE:**	18.4%
(0.05)	AR(2)	**Mean Error:**	(0.00)
		Correlation:	0.00
		Akaike:	-1.01E+05

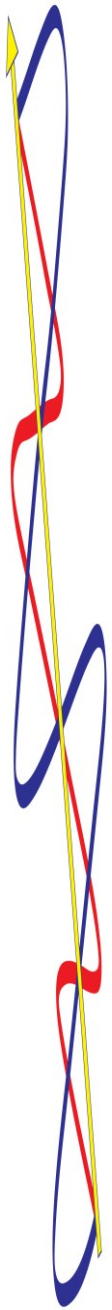
p = 2

(0.00)	alpha	**Data Points:**	12763
1.05	AR(1)	**RMSE:**	18.4%
(0.06)	AR(2)	**Mean Error:**	0.00
0.01	AR(3)	**Correlation:**	(0.00)
0.01	AR(4)	**Akaike:**	-1.01E+05

p = 4

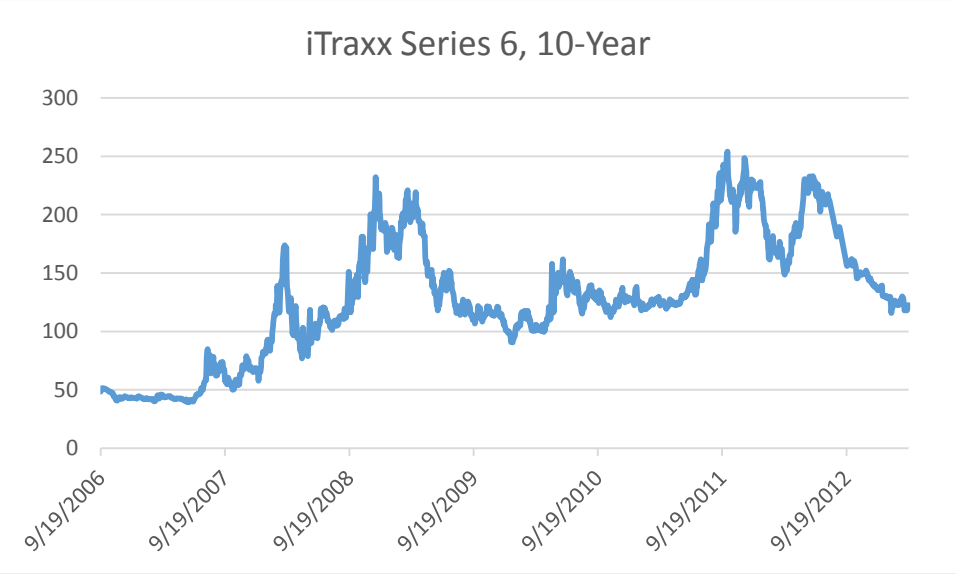
(0.00)	alpha	**Data Points:**	12763
1.05	AR(1)	**RMSE:**	18.4%
(0.06)	AR(2)	**Mean Error:**	0.00
0.01	AR(3)	**Correlation:**	0.00
(0.02)	AR(4)	**Akaike:**	-1.01E+05
0.00	AR(5)		
0.01	AR(6)		
0.04	AR(7)		
(0.03)	AR(8)		

p = 8



AR Model Output - Observations

- The $p = 1$ case validates log-normal as "good"
- Yet the $p = 2$ case is "better" since correlation between successive random variables is closer to zero
- No accuracy improvement for higher values of p – thus it's best to use $p = 1$ or $p = 2$
- No known stochastic model for $p = 2$?!



iTraxx S6 10Y		Log
date	iTraxx s6	iTraxx s6
9/19/2006	48.375	3.88
9/20/2006	50.55	3.92
9/21/2006	50.71	3.93
9/22/2006	51.371	3.94
9/25/2006	51.257	3.94
9/26/2006	50.984	3.93
9/27/2006	50.822	3.93
9/28/2006	51.087	3.93
9/29/2006	51.001	3.93
10/2/2006	50.979	3.93
10/3/2006	50.938	3.93
10/4/2006	50.527	3.92
10/5/2006	50.255	3.92

AR Model Output: the RMSE (annualized) is the iTraxx extracted volatility of 53.8%

0.02	alpha	**Data Points:**	1641
1.00	AR(1)	**RMSE:**	53.8%
		Mean Error:	0.00
		Correlation:	0.02
		Akaike:	-9.45E+03

p = 1

AR Model Output: Notice different results for different *p*

0.02	alpha	**Data Points:**	1641
1.00	AR(1)	**RMSE:**	53.8%
		Mean Error:	0.00
		Correlation:	0.02
		Akaike:	-9.45E+03

p = 1

0.02	alpha	**Data Points:**	1641
1.02	AR(1)	**RMSE:**	53.8%
(0.02)	AR(2)	**Mean Error:**	0.00
		Correlation:	0.00
		Akaike:	-9.45E+03

p = 2

0.02	alpha	**Data Points:**	1641
1.02	AR(1)	**RMSE:**	53.8%
(0.06)	AR(2)	**Mean Error:**	0.00
0.01	AR(3)	**Correlation:**	(0.00)
0.02	AR(4)	**Akaike:**	-9.45E+03

p = 4

0.02	alpha	**Data Points:**	1641
1.02	AR(1)	**RMSE:**	53.7%
(0.06)	AR(2)	**Mean Error:**	(0.00)
0.01	AR(3)	**Correlation:**	(0.00)
0.02	AR(4)	**Akaike:**	-9.44E+03
(0.05)	AR(5)		
0.09	AR(6)		
(0.05)	AR(7)		
0.02	AR(8)		

p = 8

AR Model Output - Observations

- The $p = 1$ case validates log-normal as "okay" – but we need to notice non-zero α
- The $p = 2$ case is "better" since correlation between successive random variables is closer to zero
- Small accuracy improvement for $p = 8$, but seems best to use $p = 1$ or $p = 2$
- No known stochastic model for $p = 2$?!

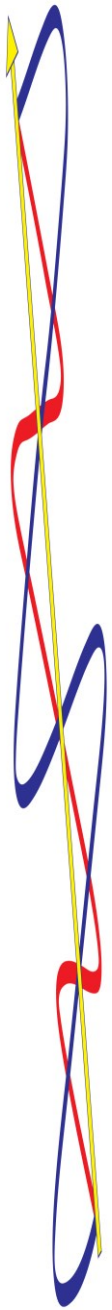
Earlier we showed these model choices:

A Normal Stochastic Model for the 10-year UST yield *y* might use standard deviation *s = 1.1%* and the random variable **ε** → N(0,1)

$$\Delta y = r\,\Delta t + s\,\varepsilon\,\sqrt{\Delta t}$$

A Log-Normal Stochastic Model for the 10-year UST yield *y* might use volatility ***σ = 18.4%*** and the random variable **ε** → N(0,1)

$$\Delta \mathrm{y} = \mu\,\mathrm{y}\,\Delta t + \sigma\,\mathrm{y}\,\varepsilon\,\sqrt{\Delta t}$$



But the AR model gives a different form when $p = 1$ and $a_1 \neq 1$:

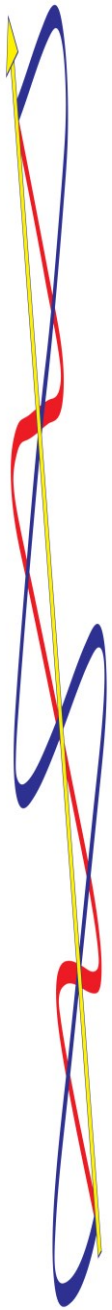
$$y_i = \alpha + \sum_{j=1}^{p} a_j y_{i-j} + \hat{s}\varepsilon = \alpha + a_1 y_{i-1} + \hat{s}\varepsilon$$

$$y_{i+1} = \alpha + a_1 y_i + \hat{s}\varepsilon$$

$$y_{i+1} - y_i = \Delta y_i = \alpha - (1 - a_1) y_i + \hat{s}\varepsilon$$

$$\Delta y_i = \omega(m - y_i)\Delta t + s\varepsilon\sqrt{\Delta t}$$

$$\omega = (1 - a_1)/\Delta t \quad \text{and} \quad m = \alpha/(1 - a_1)$$



But the AR model gives a different form when $p = 1$ and $a_1 \neq 1$:

$$\Delta y_i = \omega(m - y_i)\Delta t + s\varepsilon\sqrt{\Delta t}$$

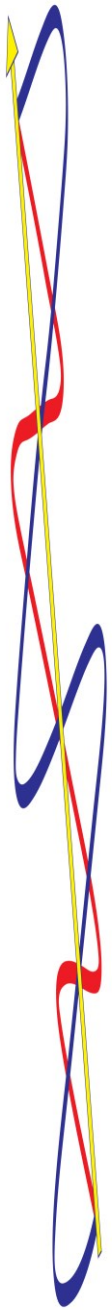
$$\omega = (1 - a_1)/\Delta t \quad \text{and} \quad m = \alpha/(1 - a_1)$$

This is Mean Reversion !

Even though we show $a_1 = 1.00$, the more precise value is $a_1 = 0.9966$ which implies mean reversion speed of 0.85 pa with Long-Term Spread m of 135 bps pa.

What's the point here ?

- We posited an initial stochastic process.
- We analyzed data with an AR framework.
- Historical analysis confirmed some aspects but also identified a new concept for iTraxx data.
- Is this new concept plausible?



"Mean Reversion" form:

$$\Delta y_i = \omega(m - y_i)\Delta t + s\varepsilon\sqrt{\Delta t}$$

Previous "Normal" form:

$$\Delta y_i = r\,\Delta t + s\,\varepsilon\,\sqrt{\Delta t}$$

Previous "Log-Normal" form:

$$\Delta y_i = \mu\, y_i\, \Delta t + \sigma\, y_i\, \varepsilon\, \sqrt{\Delta t}$$

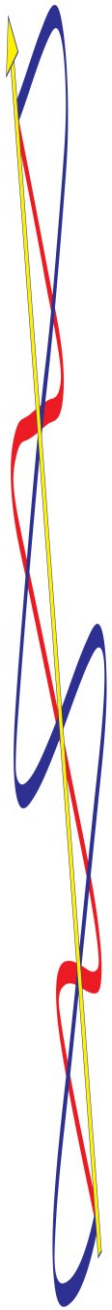
Attribute of Normal and Log-Normal models is that they have "easy" closed-form solutions for time *t*:

$$\Delta y_i = r\,\Delta t + s\,\varepsilon\,\sqrt{\Delta t}$$

$$\Rightarrow \quad y(t) = y_o + rt + s\sqrt{t}\,\varepsilon$$

$$\Delta y_i = \mu\, y_i\,\Delta t + \sigma\, y_i\,\varepsilon\,\sqrt{\Delta t}$$

$$\Rightarrow \quad y(t) = y_o \exp\left[\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma\sqrt{t}\,\varepsilon\right]$$



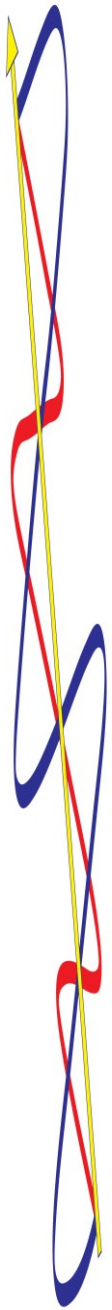
For Mean Reversion, though, "solution by inspection" doesn't work. We need stochastic differential equation (SDE) methods

$$\Delta y_i = \omega(m - y_i)\Delta t + s\varepsilon\sqrt{\Delta t}$$

becomes $$\mathrm{d}y = \omega(m - y)dt + s\,dW(t)$$

Let $\hat{y} = y - m$ → $$\mathrm{d}\hat{y} = -\omega\hat{y}dt + s\,dW(t)$$

Common trick: $$\mathrm{d}(\hat{y}\,e^{\omega t}) = s\,e^{\omega t}\,dW(t)$$



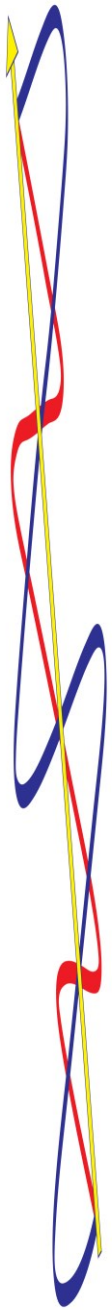
Mean Reversion .... Continued

Ornstein - Uhlenbeck

$$\mathrm{d}(\hat{y}\, e^{\omega t}) = s\, e^{\omega t}\, dW(t)$$

Integrate: $$\hat{y}\, e^{\omega t} = \hat{y}_0 + s \int_0^t e^{\omega \tau}\, dW(\tau)$$

How to do this strange integration ??



Strange Integration .....

Think of a Riemann Sum

$$\int_0^t e^{\omega\tau}\, dW(\tau) \quad \Rightarrow \quad \sum_i e^{\omega\tau_i}\, \Delta W$$

$$\Rightarrow \quad \sum_i e^{\omega\tau_i} \left(\frac{\varepsilon_i}{\sqrt{\Delta\tau}}\right) \Delta\tau$$

Just a sum of independent, normally distributed random variates

The Riemann Sum is Normally distributed

$$Mean\left\{\sum_i e^{\omega\tau_i}\left(\frac{\varepsilon_i}{\sqrt{\Delta\tau}}\right)\Delta\tau\right\} = 0$$

$$Var\left\{\sum_i e^{\omega\tau_i}\left(\frac{\varepsilon_i}{\sqrt{\Delta\tau}}\right)\Delta\tau\right\} = \sum_i e^{2\omega\tau_i}\Delta\tau$$

These results tell us what to do with the integral

Back to the strange integral ....

$$Mean\left\{\int_0^t e^{\omega\tau}\, dW(\tau)\right\} = 0$$

$$Var\left\{\int_0^t e^{\omega\tau}\, dW(\tau)\right\} = \int_0^t e^{2\omega\tau}\, d\tau$$

This last integral for the variance is straightforward .... we get

$$\frac{e^{2\omega t} - 1}{2\omega}$$

Final Result for Mean Reversion

$$\mathrm{d}y = \omega(m - y)dt + s\, dW(t)$$

has the solution

$$y(t) = y_0 e^{-\omega t} + m(1 - e^{-\omega t}) + s \left(\frac{1 - e^{-2\omega t}}{2\omega}\right)^{1/2} \varepsilon$$

Comparison

Normal:

$$y(t) = y_o + rt + s\sqrt{t}\,\varepsilon$$

Log-Normal:

$$y(t) = y_o \exp\left[\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma\sqrt{t}\,\varepsilon\right]$$

Mean Reversion:

$$y(t) = y_0 e^{-\omega t} + m(1 - e^{-\omega t}) + s\left(\frac{1 - e^{-2\omega t}}{2\omega}\right)^{1/2}\varepsilon$$

Ultimate Outcome

- "Mean Reversion" is likely the best choice when the data analysis shows $\boldsymbol{\alpha \neq 0 \; and \; a_1 \neq 1}$
- The Mean Reversion Level is taken from the data and not "made up"
- Impact on a forward projected probability distribution is huge for tail events at long time – which is further reason for caution

	Spain
date	Δ GDP %
1980	1.203
1981	-0.408
1982	1.239
1983	1.652
1984	1.698
1985	2.362
1986	3.432
1987	5.709
1988	5.285
1989	5.004
1990	3.847
1991	2.525
1992	0.851
1993	-1.314

AR Model Output: GDP data differs markedly from earlier examples – what does that mean?

0.63	alpha	**Data Points:**	33
0.72	AR(1)	**RMSE:**	1.60
		Mean Error:	0.00
		Correlation:	0.12
		Akaike:	6.81E+01

p = 1

AR Model Output: Notice different results for different *p*

0.63	alpha	**Data Points:**	33
0.72	AR(1)	**RMSE:**	1.60
		Mean Error:	0.00
		Correlation:	0.12
		Akaike:	6.81E+01

p = 1

0.93	alpha	**Data Points:**	33
0.84	AR(1)	**RMSE:**	1.55
(0.21)	AR(2)	**Mean Error:**	(0.00)
		Correlation:	0.05
		Akaike:	6.84E+01

p = 2

1.58	alpha	**Data Points:**	33
0.88	AR(1)	**RMSE:**	1.47
(0.47)	AR(2)	**Mean Error:**	(0.00)
0.48	AR(3)	**Correlation:**	(0.02)
(0.47)	AR(4)	**Akaike:**	6.99E+01

p = 4

2.15	alpha	**Data Points:**	33
0.90	AR(1)	**RMSE:**	1.44
(0.61)	AR(2)	**Mean Error:**	(0.00)
0.69	AR(3)	**Correlation:**	(0.01)
(0.84)	AR(4)	**Akaike:**	8.08E+01
0.44	AR(5)		
(0.57)	AR(6)		
0.52	AR(7)		
(0.34)	AR(8)		

p = 8

Can we trust the data?

- Different data sources and situations deserve different levels of trust
- Sovereign economic data is not high on the "trust scale" – Greece, Argentina, China, IMF, ECB
- Even with best efforts, sovereign economic data has low precision
- Behavior as p increases is odd – perhaps due to small number of data points

Conceptual Questions for Spanish GDP Data

- When building a model for sovereign default risk, how do we use historical data over a period of no default? Ignore mean reversion?
- Do we use or ignore the pre-EURO period?

Unsolved Problem re Spanish GDP Data

- With colleagues, we created a sovereign default model that treats GDP growth as a random walk (normal distribution)
- To BACKTEST the model, we chose Spain in an earlier year such as 2006 and found a problem
- The observed RMSE for GDP change for 1980-2006 is just 1.2%. But actual GDP changes in years beginning 2007, 2008, and 2009 are 2x, 4x, and 3x this 1.2% value!

Unsolved Problem re Spanish GDP Data

- Large actual GDP changes imply that the model is unsatisfactory OR that the data should be "de-weighted"
- All ideas are welcome!

Conclusion

- Time Series Data Analysis
- Treasury yields, CDS indices, Sovereign GDP
- Data analysis is necessary to understand risk positions, trading strategies, monetary and fiscal policies, *et cetera*
- But data analysis is not sufficient, it will not provide models, concepts, probability distributions, *et cetera*